+25

Midterm exam, Survival Analysis I, 2018 Spring [+ 28 points]

Name: [redacted]

- **Not only answer but also derivations**

+5 **Q1 [+8]** Let $t_i = (1650, 30, 720, 450, 510, 1110, 210, 1380, 1800, 540)$,

$\delta_i = (0, 1, 0, 1, 1, 0, 1, 1, 0, 1)$, and $x_i = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$.

$t_i(\delta_i=1)$	x_i	n_{i1}	n_{i0}	n_i	$\frac{n_{i1}}{n_i}$
30	1	5	5	10	1/2
210	0	4	5	9	4/9
450	1	4	4	8	1/2
510	1	3	4	7	3/7
540	0	2	4	6	1/3
1380	0	1	2	3	1/3

+2 (1) [+2] Calculate the log-rank statistic

$$S = \sum_{i=1}^{n} \delta_i \left(x_i - \frac{n_{i1}}{n_i} \right) = \sum_{i=1}^{n} \delta_i x_i - \sum_{i=1}^{n} \delta_i \frac{n_{i1}}{n_i}$$

$$= 3 - \frac{160}{63} = \frac{29}{63} \approx 0.46$$

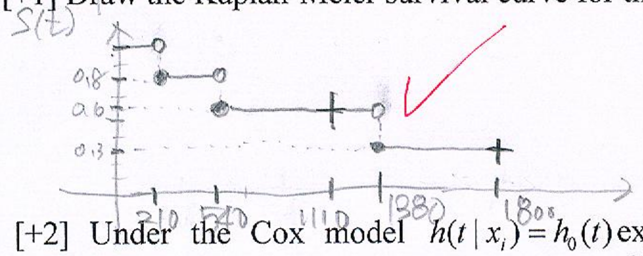
+1 (2) [+1] Calculate the variance of the log-rank statistic

$$Var(S) = \left(\frac{1}{2}\right)^2 + \frac{4}{9} \times \frac{5}{9} + \left(\frac{1}{2}\right)^2 + \frac{3}{7} \times \frac{4}{7} + \frac{1}{3} \times \frac{2}{3} + \frac{1}{3} \times \frac{2}{3} = \frac{1}{2} + \frac{3716}{3969} \approx 1.44$$

+1 (3) [+1] Calculate the Kaplan-Meier survival curve for the group of $x_i = 0$.

$t_i(\delta_i=1)$	n_i	$1/n_i$	$1-1/n_i$	$\hat{S}(t_i)$
210	5	1/5	4/5	4/5 = 0.8
540	4	1/4	3/4	$\frac{4}{5} \times \frac{3}{4} = \frac{3}{5}$ = 0.6
1380	2	1/2	1/2	$\frac{3}{5} \times \frac{1}{2} = \frac{3}{10}$ = 0.3

+1 (4) [+1] Draw the Kaplan-Meier survival curve for the group of $x_i = 0$.



Since

$$\sum_{l \in R_i} e^{\beta x_l} = \sum_{\substack{l \in R_i \\ x_l = 0}} 1 + \sum_{\substack{l \in R_i \\ x_l = 1}} e^{\beta} = n_{i0} + n_{i1} e^{\beta}$$

$$\sum_{l \in R_i} x_l e^{\beta x_l} = \sum_{\substack{l \in R_i \\ x_l = 0}} 0 + \sum_{\substack{l \in R_i \\ x_l = 1}} e^{\beta} = n_{i1} e^{\beta}$$

(5) [+2] Under the Cox model $h(t \mid x_i) = h_0(t)\exp(\beta x_i)$, derive the fixed point iteration algorithm.

The partial likelihood is $L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta x_i}}{\sum_{l \in R_i} e^{\beta x_l}} \right)^{\delta_i}$, where $R_i = \{ l : t_l \geq t_i \}$ is the risk set

The log partial likelihood is $l(\beta) = \sum_{i=1}^{n} \delta_i \beta x_i - \sum_{i=1}^{n} \delta_i \log\left(\sum_{l \in R_i} e^{\beta x_l} \right)$

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i x_i - \sum_{i=1}^{n} \delta_i \frac{\sum_{l \in R_i} x_l e^{\beta x_l}}{\sum_{l \in R_i} e^{\beta x_l}}$$

$$= \sum_{i=1}^{n} \delta_i x_i - \sum_{i=1}^{n} \delta_i \frac{n_{i1} e^{\beta}}{n_{i0} + n_{i1} e^{\beta}} \overset{set}{=} 0 \Rightarrow e^{\beta} = \frac{\sum_{i=1}^{n} \delta_i x_i}{\sum_{i=1}^{n} \delta_i \frac{n_{i1}}{n_{i0} + n_{i1} e^{\beta}}}$$

Algorithm

Step 1: Choose initial $\beta^{(0)} = 0$.

Step 2: Repeat

$$\beta^{(k+1)} = \log\left(\frac{\sum_{i=1}^{n} \delta_i x_i}{\sum_{i=1}^{n} \delta_i \frac{n_{i1}}{n_{i0} + n_{i1} e^{\beta^{(k)}}}} \right)$$

* If $|\beta^{(k+1)} - \beta^{(k)}| < \varepsilon$, for some small ε, then stop and set $\hat{\beta} = \beta^{(k)}$.

(6) [+1] Calculate the first step of the fixed-point iteration

$$\exp(\beta^{(1)}) = \frac{\sum_{i=1}^{n} \delta_i x_i}{\sum_{i=1}^{n} \delta_i \frac{n_{i1}}{n_{i0} + n_{i1}}} = \frac{3}{\frac{160}{63}} = \frac{189}{160} \approx 1.18$$

Please check if Algorithm works. If so, I can add +3 (Report)

+6 **Q2 [+6]** The hazard function follow the Cox model
$h(t \mid x_1, x_2, x_3) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3)$, where the gene expression values are

$$x_1 = \begin{cases} 1 & \text{high value of NCOA3} \\ 0 & \text{low value of NCOA3} \end{cases}, \qquad \beta_1 = 0.237$$

$$x_2 = \begin{cases} 1 & \text{high value of TEAD1} \\ 0 & \text{low value of TEAD1} \end{cases}, \qquad \beta_2 = 0.223$$

$$x_3 = \begin{cases} 1 & \text{high value of YWHAB} \\ 0 & \text{low value of YWHAB} \end{cases}, \qquad \beta_3 = 0.263$$

Compute the relative risk (RR).

+1 1) [+1] RR of (all three genes in high value) vs. (all three genes in low value).
RR=exp(0.723) ✓

+1 2) [+1] RR of (all three genes in high value) vs. (only NCOA3 in high value).
RR=exp(0.486) ✓

+1 3) [+1] RR of (only NCOA3 in high value) vs. (only YWHAB in high value).
RR=exp(-0.026) ✓

+3 4) [+3] **All** RRs under different combinations risk factors (vs. the baseline risk).
Make a table by sorting the RRs (from highest to lowest).

Write →

Order	RR	NCOA3	TEAD1	YWHAB
1	exp(0.723)	High	High	High
2	exp(0.5)	High	Low	High
3	exp(0.486)	Low	High	High
4	exp(0.46)	High	High	Low
5	exp(0.263)	Low	Low	High.
6	exp(0.237)	High	Low	Low
7	exp(0.223)	Low	High	Low
8	exp(0)	Low	Low	Low

+10

Q3 [+10] Let (t_i, δ_i), $i = 1, \ldots, n$, be survival data. Let $m = \sum_{i=1}^{n} \delta_i$, $m^* = \sum_{i=1}^{n} (1-\delta_i)$,

$S = \sum_{i=1}^{n} \delta_i t_i$, and $S^* = \sum_{i=1}^{n} (1-\delta_i) t_i$. Let $\Pr(T > t, U > u) = [\exp(\lambda t) + \exp(\mu u) - 1]^{-1}$.

+2 (1) [+2] Derive the cause-specific hazard functions $h_T^{\#}(t)$ and $h_U^{\#}(t)$.

$$h_T^{\#}(t) = -\frac{\partial}{\partial x} \log P(T > x, U > t)\Big|_{x=t} = \frac{-\frac{\partial}{\partial x} P(T > x, U > t)}{P(T > x, U > t)}\Big|_{x=t} = \frac{\lambda e^{\lambda t}}{(e^{\lambda t} + e^{\mu t} - 1)}$$

Similarly, $h_U^{\#}(t) = \dfrac{\mu e^{\mu t}}{(e^{\lambda t} + e^{\mu t} - 1)}$

+2 (2) [+2] Derive the log-likelihood function $\ell(\lambda, \mu)$. (simplify the answer)

The likelihood function is $L(\lambda, \mu) = \prod_{i=1}^{n} h_T(t_i)^{\delta_i} h_U(t_i)^{1-\delta_i} \cdot P(T > t_i, U > t_i)$

Then the log-likelihood function is

$$\ell(\lambda, \mu) = \sum_{i=1}^{n} \delta_i \log h_T^{\#}(t_i) + \sum_{i=1}^{n} (1-\delta_i) \log h_U^{\#}(t_i) + \sum_{i=1}^{n} \log P(T > t_i, U > t_i)$$

$$= m \log \lambda + S\lambda + m^* \log \mu + S^* \mu - 2\sum_{i=1}^{n} \log(e^{\lambda t_i} + e^{\mu t_i} - 1).$$

+2 (3) [+2] Write the score equations s.t. $\begin{cases} \lambda = f(\lambda, \mu) \\ \mu = g(\lambda, \mu) \end{cases}$ for functions f and g.

$$\frac{\partial \ell(\lambda, \mu)}{\partial \lambda} = \frac{m}{\lambda} + S - 2\sum_{i=1}^{n} \frac{t_i e^{\lambda t_i}}{e^{\lambda t_i} + e^{\mu t_i} - 1} \overset{\text{set}}{=} 0 \Rightarrow \lambda = \left(\frac{2}{m} \sum_{i=1}^{n} \frac{t_i e^{\lambda t_i}}{e^{\lambda t_i} + e^{\mu t_i} - 1} - \frac{S}{m} \right)^{-1}$$

$$\frac{\partial \ell(\lambda, \mu)}{\partial \mu} = \frac{m^*}{\mu} + S^* - 2\sum_{i=1}^{n} \frac{t_i e^{\mu t_i}}{e^{\lambda t_i} + e^{\mu t_i} - 1} \overset{\text{set}}{=} 0 \Rightarrow \mu = \left(\frac{2}{m^*} \sum_{i=1}^{n} \frac{t_i e^{\mu t_i}}{e^{\lambda t_i} + e^{\mu t_i} - 1} - \frac{S^*}{m^*} \right)^{-1}$$

+2 (4) [+2] Write the fixed-point iteration algorithm by the above results.

Algorithm:

Step 1: Choose initial parameters $\lambda^{(0)}, \beta^{(0)}$

Step 2: Update $\lambda^{(k+1)} = \left(\frac{2}{m} \sum_{i=1}^{n} \frac{t_i e^{\lambda^{(k)} t_i}}{e^{\lambda^{(k)} t_i} + e^{\mu^{(k)} t_i} - 1} - \frac{S}{m} \right)^{-1}$

Step 3: Update $\mu^{(k+1)} = \left(\frac{2}{m^*} \sum_{i=1}^{n} \frac{t_i e^{\mu^{(k)} t_i}}{e^{\lambda^{(k+1)} t_i} + e^{\mu^{(k)} t_i} - 1} - \frac{S^*}{m^*} \right)^{-1}$

Step 4: Repeat Step 2 – Step 3 as $k = 0, 1, 2, \ldots$

* If $\max\{|\lambda^{(k+1)} - \lambda^{(k)}|, |\mu^{(k+1)} - \mu^{(k)}|\} < \varepsilon$ for some small ε, then stop and set $\hat{\lambda} = \lambda^{(k)}, \hat{\mu} = \mu^{(k)}$

+2 (5) [+2] Assume $\mu = 1$ is known. Write the Newton-Rapson algorithm for λ.

If $\mu = 1$, we have

$$\ell(\lambda) = m \log \lambda + S\lambda + S^* - 2\sum_{i=1}^{n} \log(e^{\lambda t_i} + e^{t_i} - 1)$$

$$S(\lambda) = \frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{m}{\lambda} + S - 2\sum_{i=1}^{n} \frac{t_i e^{\lambda t_i}}{e^{\lambda t_i} + e^{t_i} - 1}$$

$$H(\lambda) = \frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} = -\frac{m}{\lambda^2} - 2\sum_{i=1}^{n} \left\{ \frac{t_i^2 e^{\lambda t_i}}{e^{\lambda t_i} + e^{t_i} - 1} - \frac{t_i^2 e^{2\lambda t_i}}{(e^{\lambda t_i} + e^{t_i} - 1)^2} \right\}$$

Algorithm:

Step 1: Choose initial parameter $\lambda^{(0)}$

Step 2: Repeat the Newton-Raphson iterations

$$\lambda^{(k+1)} = \lambda^{(k)} - H^{-1}(\lambda^{(k)}) S(\lambda^{(k)})$$

as $k = 0, 1, 2, \ldots$

* If $|\lambda^{(k+1)} - \lambda^{(k)}| < \varepsilon$ for some small ε, then stop and set $\hat{\lambda} = \lambda^{(k)}$.

+4 **Q4 [+4]** Let (t_i, δ_i), $i = 1, \ldots, n$, be survival data.

Derive the Kaplan-Meier estimator $\hat{S}(t)$ under the following assumptions:

(A1) $S(t) = \Pr(T > t)$ is a step function with jumps at death times.

(A2) There are no ties in the data.

(A3) Censoring time is independent of survival time.

In the derivation, explain how to use (A1)-(A3).

Suppose $0 = t_0 < t_1 < t_2 < \cdots < t_n$.. are distinct death times by (A2),

Then,
$$P(T > t_n) = P(T > t_n \mid T > t_{n-1}) P(T > t_{n-1})$$
$$= P(T > t_n \mid T > t_{n-1}) P(T > t_{n-1} \mid T > t_{n-2}) P(T > t_{n-2})$$
$$= \cdots = \prod_{i=1}^{n} P(T > t_i \mid T > t_{i-1})$$

By (A1), we have
$$\prod_{i=1}^{n} P(T > t_i \mid T > t_{i-1}) = \prod_{i=1}^{n} P(T > t_i \mid T \geq t_i)$$

Then.
$$\prod_{i=1}^{n} P(T > t_i \mid T \geq t_i) = \prod_{i=1}^{n} \left\{ 1 - P(T \leq t_i \mid T \geq t_i) \right\}$$
$$= \prod_{i=1}^{n} \left\{ 1 - \frac{P(T = t_i)}{P(T \geq t_i)} \right\}$$

Suppose U is the censoring time, by (A3), we obtain.

$$\frac{P(T = t_i)}{P(T \geq t_i)} = \frac{P(T = t_i, U \geq t_i)}{P(T \geq t_i, U \geq t_i)} = \frac{P(\min(T, U) = t_i, T \leq U)}{P(\min(T, U) \geq t_i)}.$$

We estimate $P(\min(T, U) = t_i, T \leq U)$ by $\frac{1}{n} \sum_{l=1}^{n} I(t_l = t_i, \delta_l = 1) = \frac{1}{n}$

due to no ties assumption (A2)

We estimate $P(\min(T, U) \geq t_i)$ by $\frac{1}{n} \sum_{l=1}^{n} I(t_l \geq t_i) \equiv \frac{n_i}{n}$

Finally, we estimate $S(t)$ by

$$\hat{S}(t) = \prod_{\substack{t_i \leq t \\ \delta_i = 1}} \left\{ 1 - \frac{1}{n_i} \right\}, \quad t \in [0, \max_{i=1,\ldots,n}(t_i)]$$